Developing a Protocol for Annotating Sets of Closely-related Genomes: Example Genus *Brucella*

Terrence Disz[1], Janaka N. Edirisinghe[1,5], José P. Faria[1,4], Anna Hausmann[2], Christopher S. Henry[1,5], Robert Olson[1], Ross A. Overbeek[1,2], Gordon D. Pusch[2], Maulik Shukla[3], Veronika Vonstein[1,2] and Alice R. Wattam[3]

[1]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA

[2]Fellowship for Interpretation of Genomes, Burr Ridge, IL, USA

[3]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

[4]IBB-Institute for Biotechnology and Bioengineering/Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[5]Computation Institute, University of Chicago, Chicago, Illinois, USA

Abstract

Here we report on a protocol for the improvement of initial automated annotations for sets of closely-related genomes across a genus. This is an enabling step for the future creation of accurate, predictive models that will support reasoning about differences in phenotype, media requirements etc. This protocol includes the generation of protein families for fifteen publicly available genomes from the genus *Brucella*, their manual curation and metrics to evaluate the improvements.

Introduction

Since the first bacterial genome was sequenced in 1995 (Fleischmann et al. 1995) the number of genomic sequences has grown exponentially (Lagesen et al. 2010). It has become possible to routinely generate fairly accurate annotations (Aziz et al. 2008), and initial metabolic models with minimal effort (Henry et al. 2010) though the creation of accurate, predictive models requires a substantial investment in phenotypic data (e.g., BioLog or RNAseq data) and iterative reconciliation with the models.

The *Brucella* are a group of intracellular facultative bacteria that cause disease in humans and many animals of economic importance. Phylogenetically they have distinct radiations that result in well-supported branching patterns and yet are well

conserved with little divergence (Wattam et al. 2009), making them ideal for an initial foray into resolving annotation discrepancies at the genus level. Additionally, PATRIC, the all bacterial bioinformatics resource center (Gillespie et al. 2011) has strong collaborations with the *Brucella* community.

The quality of initial metabolic network reconstructions and predictive metabolic models depends on the quality and consistency of the annotations from which they were generated. Isofunctional homologs need to have the same annotations, so that they can be connected to the same reactions in the models. If one attempts to compare initial metabolic reconstructions for distinct organisms utilizing automated annotations, a significant number of discrepancies in the resulting models are specious.

Methods and Results

Brucella Genomes.

Fifteen publicly available genomes from the genus *Brucella* were chosen to represent all of the clades within the genus (Wattam et al. 2012) (Table 1).

Generating Mobile Element Families.

Within the fifteen genomes we identified repeat regions longer then 100 bp and with DNA identity greater than, or equal to 90%. Such regions (other than multiple rRNA islands) often reflect events involving mobile elements. We formed protein families for the protein-encoding genes of those regions, and called this set of families *the mobile element protein families.* There are 50 mobile element protein

families containing 410 proteins. The functions of the proteins in the sets are considered neither reliable nor consistent. We did not attempt to correct them, since the proteins, while often related to virulence, are seldom included in metabolic models.

Generating Non-Mobile Element Families

Two proteins were placed in the same protein family if they were bi-directional best hits with greater than 50% identify over 80% length of each of the two proteins, and the genes occurred within a conserved context. We considered the context of the matched pairs to be conserved, if there were at least 3 pairs of bi-directional best hits co-occurring within a 10 Kb region. We produced 5,038 families (with two or more proteins) containing a total of 52,626 proteins. From these families we generated *core protein families*. Each of these families contained at most one protein from each genome, and 80% of the genomes were represented in the family. All non-mobile element protein families were included in the supplementary materials.

Removing Annotation Inconsistencies to Support Metabolic Network Reconstruction

Our first goal was to improve the consistency of annotations of the representative genomes. We constructed tools for detecting and curating families that contained inconsistent annotations. A total of 398 families containing 4,848 proteins were manually curated.

We defined two metrics to measure progress.

The first:

> *Given two proteins from the same protein family (i.e., from one of the 5,038 families we constructed), what is the probability that they have been assigned precisely the same function?*

We computed this property and compared our annotations to other public annotation recourses (Table 2).

The second:

> *How many Brucella-universal-reactions have been assigned to each genome?*

By universal reactions we mean the reactions that are present in all *Brucella* genomes. We chose this second metric to demonstrate that improvements in annotations lead to improvements in the metabolic reconstructions.

Metabolic Network Reconstructions

Three sets of metabolic reconstructions were generated for the fifteen *Brucella* genomes (Supplementary Material S1 and S2). The initial set was built with the annotations provided by RAST, while a second set of reconstructions was generated after the manual curation of inconsistent protein families. Programs were written

to build, compare and analyze the metabolic reconstruction networks. The set of non-universal reactions was analyzed in detail and lead to an additional round of annotation improvements, which in turn resulted in an improved the third set of metabolic reconstructions.

The initial set of metabolic reconstructions contained 1011 *Brucella*-universal-reactions. In the second set of reconstructions, the number of *Brucella*-universal-reactions increased to 1016 and reached a total of 1047 *Brucella*-universal-reactions in the third set.

The 86 non-universal reactions from the second set of reconstructions were analyzed manually and revealed problems with the assertion or omission of some of those reactions in certain genomes (Supplementary Material S3). We verified the absence of 39 reactions from the set of genomes and identified 24 cases of *Brucella*-universal-reactions that had been not identified in the first rounds of metabolic reconstructions. The leading cause for the omission of reactions was insufficient sequencing quality (frame shifts, incomplete ORFs at the end of contigs or stretches of low quality sequence leading to gene call errors). We found 16 cases of outdated functional roles and errors in the reaction database (labeled as "Functional role ambiguities" in Table S3). We verified five unique non-universal reactions in the *Brucella inopinata* BO1 and *Brucella inopinata*-like BO2 strains. Those reactions are involved in rhamnose-containing glycan synthesis. Additionally we proposed a candidate protein for the missing step in the diaminopimelate pathway (DAP) of leucine biosynthesis. Lastly, we found one gene fusion, that was not properly

identified by RAST and two minor errors caused by the imperfection of the family calling algorithm. The discovery and improvement of those errors was made possible by the iteration of manual curation of annotations followed by the analysis of the resulting metabolic reconstruction. All *Brucella* non-universal reactions for each genome are provided in the Supplementary Material Table S4 and related functional roles on Table S5.

Discussion

We have produced an accurate and consistent collection of annotations and initial estimates of the metabolic network for the genus *Brucella*. By manual curation of 398 protein families whose members had different annotations for isofunctional homologs, we lowered the percentage of inconsistently annotated pairs from 0.6% to 0.3%. Those improvements lead to changes in the metabolic reconstructions, generating a larger set of universal reactions and highlighting real metabolic differences for certain organisms. An analysis of the non-universal reactions after the curation of the protein families, further improved the consistency of the annotations and subsequently the metabolic reconstructions. Annotation errors due to sequencing errors were identified, along with pathway fixes and insights of unique biology in several genomes. This analysis improved both annotations and the ModelSEED reaction database by flagging ambiguities in current functional roles. We also show that metabolic reconstructions can be used as a measure of annotation consistency.

The comparison to other publicly available annotations for the same genomes illustrates how difficult the creation and especially comparison of any metabolic models might become, if any of those sources would be used without similar reconciliation steps, as we have worked out in this report.

The resulting collection of protein families, annotations, and reactions will be used to substantially improve our ability to provide accurate, automated annotations for *Brucellaceae via* the RAST system. We have updated our FIGfam collection with the improved protein families. In addition, all changes that we report have been propagated to the RAST server, the PubSEED (Aziz et al. 2012), and to PATRIC(Gillespie et al. 2011). This makes the data visible to the research community and supports further annotation efforts. This in turn will support efforts to create accurate predictive metabolic models (Devoid et al. 2013). Furthermore, the improved annotations, initially performed on 15 genomes, will propagate across all available genomes, which number 185 as of June 2013 (patricbrc.org).

This collection of data represents the key to accurate, high-throughput annotation and metabolic reconstruction of *Brucella* genomes. With this proof of concept, we plan to use this methodology to improve annotations of other conserved genera and extend it to less conserved phylogenetic groups.

Supplementary Material:

The supplementary material is available via the PATRIC website at:
http://enews.patricbrc.org/annotation_protocol_brucella/

Acknowledgements

We thank Jean Jacques Letesson, Maite Iriarte, Stephan Köhler and David O'Callaghan for their input on improving specific annotations, and Jim Davis for his helpful critique of the manuscript. This project has been funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C, awarded to BW Sobral. J.P.F. acknowledges funding from [FRH/BD/70824/2010] of the FCT (Portuguese Foundation for Science and Technology) PhD scholarship. This work was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357.

Table 1. *Brucella* genomes used in the analysis with their PubSEED and PATRIC identifiers, sizes, number of contigs, and number of CDSs

Genome Name	PubSEED ID	PATRIC Genome ID	Genome Size (bp)	Number of Contigs	Number of CDSs
Brucella abortus bv. 1 str. 9-941	262698.4	15061	3286445	2	3413
Brucella canis ATCC 23365	483179.4	25663	3312769	2	3394
Brucella ceti str. Cudo	595497.3	28239	3389269	7	3578
Brucella ceti M13/05/1	520460.3	83544	3337230	22	3367
Brucella melitensis bv. 1 str. 16M	224914.11	92729	3294931	2	3446
Brucella microti CCM 4915	568815.3	92249	3294931	2	3374
Brucella neotomae 5K33	520456.3	114381	3329623	11	3383
Brucella ovis ATCC 25840	444178.3	136990	3275590	2	3499
Brucella pinnipedialis M292/94/1	520462.3	74143	3373519	15	3356
Brucella sp. 83/13	520449.3	75385	3153851	20	3152
Brucella inopinata BO1	470735.4	109945	3366774	55	3361
Brucella inopinata-like BO2	693750.4	146994	3305941	174	3276
Brucella sp. NVSL 07-0026	520448.3	103899	3297137	17	3442
Brucella suis 1330	204722.5	107850	3315175	2	3402
Brucella suis bv. 5 str. 513	520489.3	73489	3323676	19	3316

Table 2. Consistency of annotations across different resources.
Annotations for the same 15 genomes were collected from RefSeq (Pruitt et al. 2007), the manually curated proteins from the UniProt Knowledgebase (UniProtKB)(The Universal Protein Resource (UniProt) in 2010 2010), the Translated EMBL Nucleotide Sequence Data Library (Tremblay et al. ; Boeckmann et al. 2003), the Integrated Microbial Genomes (IMG) system (Markowitz et al. 2012) and the SEED (Overbeek et al. 2005). In each case we computed the number of pairs of identical sequences to two representatives from one of our protein families. Consistency was measured based on an exact match between the functional names associated with the sequences within each source. It is worth noting that the annotations in SEED and those produced by RAST are significantly more consistent (but not necessarily better) than those from the other sources.

Source	Number of Pairs	Inconsistently Annotated	Percent of Pairs Inconsistently Annotated
RefSeq	562597217	383808122	68.2
IMG	101525838	52434525	51.6
TREMBL	112735194	46284849	41.1
SwissProt	803819	42429	5.3
SEED	271622566	9056551	3.3
Original RAST Output	16349603	102097	0.6
RAST After Manual Curation	16349603	47504	0.3

References

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. doi:10.1186/1471-2164-9-75

Aziz RK, Devoid S, Disz T, Edwards RA, Henry CS, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Stevens RL, Vonstein V, Xia F (2012) SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. PLoS One 7 (10):e48053. doi:10.1371/journal.pone.0048053

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31 (1):365-370

Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C (2013) Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol Biol 985:17-45. doi:10.1007/978-1-62703-299-5_2

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269 (5223):496-512

Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun 79 (11):4286-4298. doi:10.1128/IAI.00207-11

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28 (9):977-982. doi:10.1038/nbt.1672

Lagesen K, Ussery DW, Wassenaar TM (2010) Genome update: the 1000th genome--a cautionary tale. Microbiology 156 (Pt 3):603-608. doi:10.1099/mic.0.038257-0

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40 (Database issue):D115-122. doi:10.1093/nar/gkr1044

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L,

Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33 (17):5691-5702. doi:10.1093/nar/gki866

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35 (Database issue):D61-65. doi:gkl842 [pii] 10.1093/nar/gkl842

Tremblay PL, Summers ZM, Glaven RH, Nevin KP, Zengler K, Barrett CL, Qiu Y, Palsson BO, Lovley DR A c-type cytochrome and a transcriptional regulator responsible for enhanced extracellular electron transfer in Geobacter sulfurreducens revealed by adaptive evolution. Environ Microbiol. doi:EMI2302 [pii] 10.1111/j.1462-2920.2010.02302.x

The Universal Protein Resource (UniProt) in 2010 (2010). Nucleic Acids Res 38 (Database issue):D142-148. doi:10.1093/nar/gkp846

Wattam AR, Inzana TJ, Williams KP, Mane SP, Shukla M, Almeida NF, Dickerman AW, Mason S, Moriyon I, O'Callaghan D, Whatmore AM, Sobral BW, Tiller RV, Hoffmaster AR, Frace MA, De Castro C, Molinaro A, Boyle SM, De BK, Setubal JC (2012) Comparative genomics of early-diverging Brucella strains reveals a novel lipopolysaccharide biosynthesis pathway. MBio 3 (5):e00246-00211. doi:10.1128/mBio.00388-12

Wattam AR, Williams KP, Snyder EE, Almeida NF, Jr., Shukla M, Dickerman AW, Crasta OR, Kenyon R, Lu J, Shallom JM, Yoo H, Ficht TA, Tsolis RM, Munk C, Tapia R, Han CS, Detter JC, Bruce D, Brettin TS, Sobral BW, Boyle SM, Setubal JC (2009) Analysis of ten Brucella genomes reveals evidence for horizontal gene transfer despite a preferred intracellular lifestyle. J Bacteriol 191 (11):3569-3579. doi:10.1128/JB.01767-08